M249/Specimen

Module Examination

Practical modern statistics

Time allowed: 3 hours

There are FOUR parts to this examination. Each part carries 28 marks. From each part, a maximum of 25 marks will count towards your overall score for the examination. So you do not need to answer all questions in each part to obtain full marks.

You are advised to spend about 45 minutes on each part. A good strategy is to attempt about 25 marks' worth of questions from each of the four sections before attempting the remaining questions in any one part.

Your answers should be written *in the spaces provided* inside this examination paper. If you need extra space you may continue your working in a supplementary answer book.

At the end of the examination

Check that you have written your personal identifier and examination number on each answer book used. **Failure to do so will mean that your work cannot be identified.**

Attach your desk record on top of this examination paper using the round paper fastener.

Examination No.								
Personal Identifier								

If you have used a supplementary answer book, please attach this to the back of your examination paper using the same round paper fastener.

For Examiner's use only:

Part	1	2	3	4	Total
Marks					

Copyright © 2012 The Open University **WEB 03478 5**

5.1

PART 1

Questions 1 to 9 relate to Book 1 Medical statistics.

Write your answers in the spaces provided.

Question 1

A study is to be undertaken to investigate the association between childhood asthma and academic performance. The study will involve two groups of children: one group of children with asthma, and one group without asthma. The children will be enrolled into the study at age 5, and the proportions of children in each group who achieve a given academic standard by age 16 will be compared.

(a) Say whether this study is a cohort study or a case-control study, and give a reason for your answer. [*2*]

(b) Suppose that it is proposed that the academic performance of the children with asthma is followed closely during the 11-year follow-up period, but that only a single attempt to follow up the children without asthma is done at age 16. What bias will this follow-up strategy induce? Suggest how this bias might affect the results. [*2*]

Question 2

A case-control study of the role of heredity in Graves' disease (a disease of the thyroid gland) included 210 cases and 358 controls. An individual is exposed if he or she has more than one relative with any type of thyroid disease, and unexposed otherwise. The data are in Table 1.

Table 1

Exposure	Cases	Controls
Exposed	65	43
Unexposed	145	315
Total	210	358

(a) Estimate the odds ratio for the association between Graves' disease and having more than one relative with thyroid disease. [*1*]

(b) Obtain a 95% confidence interval for the odds ratio. [*2*]

(c) Summarize and interpret your results. [*2*]

Question 3

A study was undertaken of the association between consumption of cruciferous vegetables (such as cabbage and broccoli) and lung cancer. Participants were classified into three groups according to the amount of cruciferous vegetables they consumed: low, medium and high. The data from the study are in Table 2.

Table 2

Diet type	Lung cancer	No lung cancer	Total
Low	327	250	577
Medium	677	754	1431
High	1137	1164	2301
Total	2141	2168	4309

Suppose that interest centres on testing the null hypothesis of no association between diet type and lung cancer. The test yields $\chi^2 = 14.57$. Give the degrees of freedom for this test and a range of values for the p value. Hence interpret this result. [*2*]

Question 4

A 1–1 matched case-control study was undertaken among psychiatric patients of the association between reduction in care and suicide. There were 149 cases of suicide, individually matched to surviving psychiatric patients. The exposure is reduction in care at the last appointment. The data are in Table 3.

Table 3

		Controls	
		Care reduced	Care not reduced
Cases	Care reduced	16	50
	Care not reduced	23	60

(a) Obtain the Mantel–Haenszel odds ratio for the association between reduction in care and suicide among psychiatric patients. [*1*]

(b) The controls were matched to the cases on the basis of age. Suggest a reason why the researchers chose to do this. [*1*]

Question 5

A study was undertaken in women of the association between experiencing abuse in childhood and suffering from serious depression later in life. The severity of the abuse was classified on an increasing scale as None, Mild, Moderate or Severe. The dose-specific odds ratios relative to the None category are shown in Table 4.

Table 4

Severity of abuse	Severe	Moderate	Mild	None
Dose-specific odds ratio	4.50	3.59	2.12	1

The value of the test statistic in a chi-squared test for no linear trend between severity of abuse and serious depression later in life is 61.85, and the significance probability is $p < 0.0005$. Interpret the results in Table 4, and comment on the presence or otherwise of a dose-response relationship. [*2*]

Question 6

A study of the association between smoking and Alzheimer's disease was undertaken. The data were split into two groups, A and B, according to genetic factors. The data, stratified by group, are shown in Table 5.

Table 5

Group A

	Alzheimer's disease	No Alzheimer's disease
Smoked	20	120
Did not smoke	19	119

Group B

	Alzheimer's disease	No Alzheimer's disease
Smoked	5	12
Did not smoke	41	101

(a) Obtain the stratum-specific odds ratios for groups A and B, and the odds ratio for the aggregated data. [*3*]

(b) Comment on the role played by genetic group in the association between smoking and Alzheimer's disease. [*2*]

Question 7

A randomized controlled trial is planned of the use of magnetic imaging techniques to reduce the numbers of Caesarean sections in breech births. The trial will comprise two groups: a control group and an intervention group. In the control group, the anticipated proportion of births by Caesarean section is 0.35. A reduction of 0.15 in this proportion is regarded as being of practical significance. The power required is 80%, and the significance level is 5%.

Calculate the sample size required in each group. [*3*]

Question 8

A randomized controlled trial of atorvastatin versus placebo was undertaken in patients with rheumatoid arthritis. Participants whose condition improved after six months were classified as responders. Figure 1 shows the flow chart for the trial.

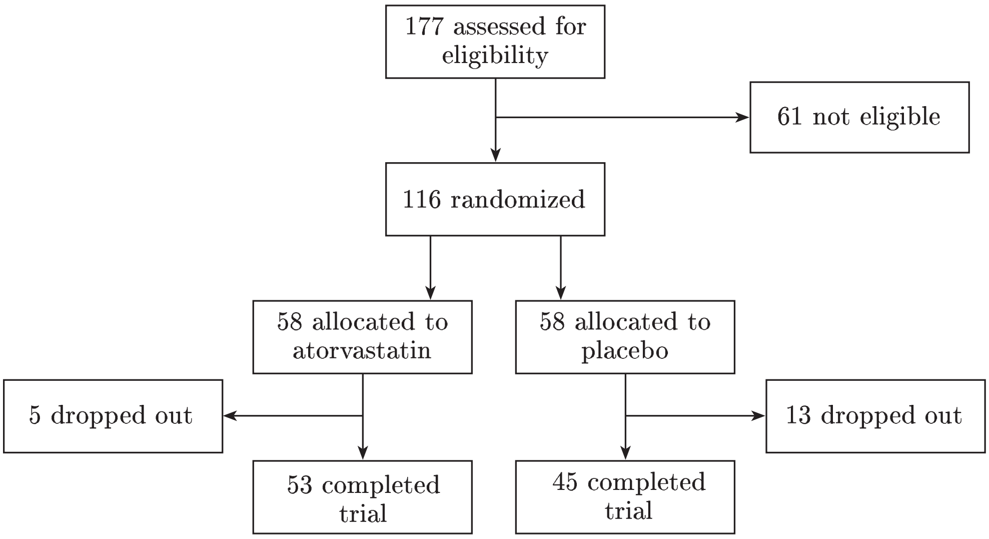

Figure 1

Of the 53 participants in the atorvastatin group who completed the trial, 18 had improved by the end of the trial. Of the 45 in the control group who completed the trial, 6 had improved. Individuals who dropped out were classified as non-responders.

Obtain the relative risk of responding to treatment in the atorvastatin group compared to the control group, for both an intention-to-treat analysis and a per-protocol analysis. [*2*]

Question 9

Seven studies of the association between long-term use of oral contraceptives (compared to never using oral contraceptives) and cervical cancer were included in a meta-analysis. The forest plot is shown in Figure 2.

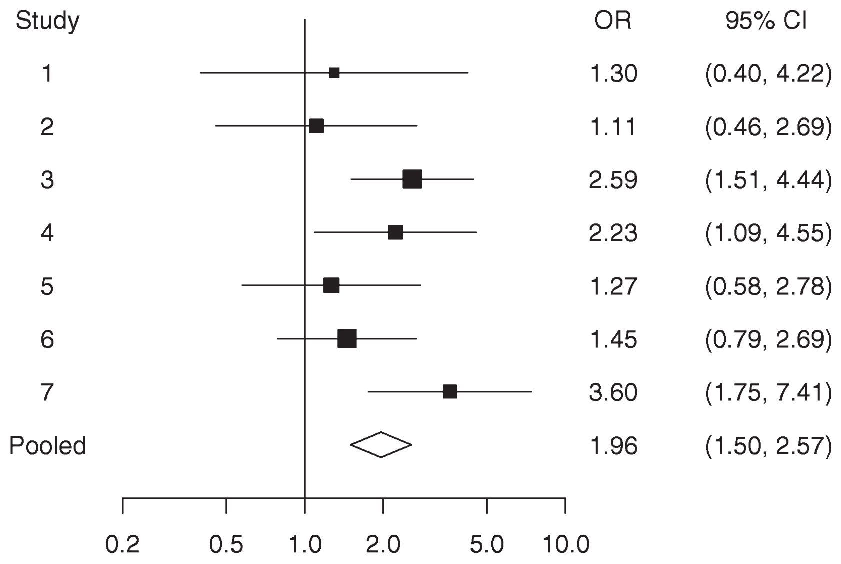

Figure 2

(a) Identify the two studies that contributed most to the pooled odds ratio. Give a reason for your answer. [*1*]

(b) Tarone's test for homogeneity of the odds ratio yields $\chi^2 = 8.14$. State the degrees of freedom for the test, and obtain an approximate p value. What do you conclude? [*2*]

For Examiner's use only:

Question No.	1	2	3	4	5	6	7	8	9	Total for Part 1
Mark										

PART 2

Questions 10 to 18 relate to Book 2 Time series.

Write your answers in the spaces provided.

Question 10

Figure 3(a) shows the time plot of a seasonally adjusted time series. Figures 3(b), 3(c) and 3(d) show three simple moving averages of the time series shown in Figure 3(a), drawn using the same scale.

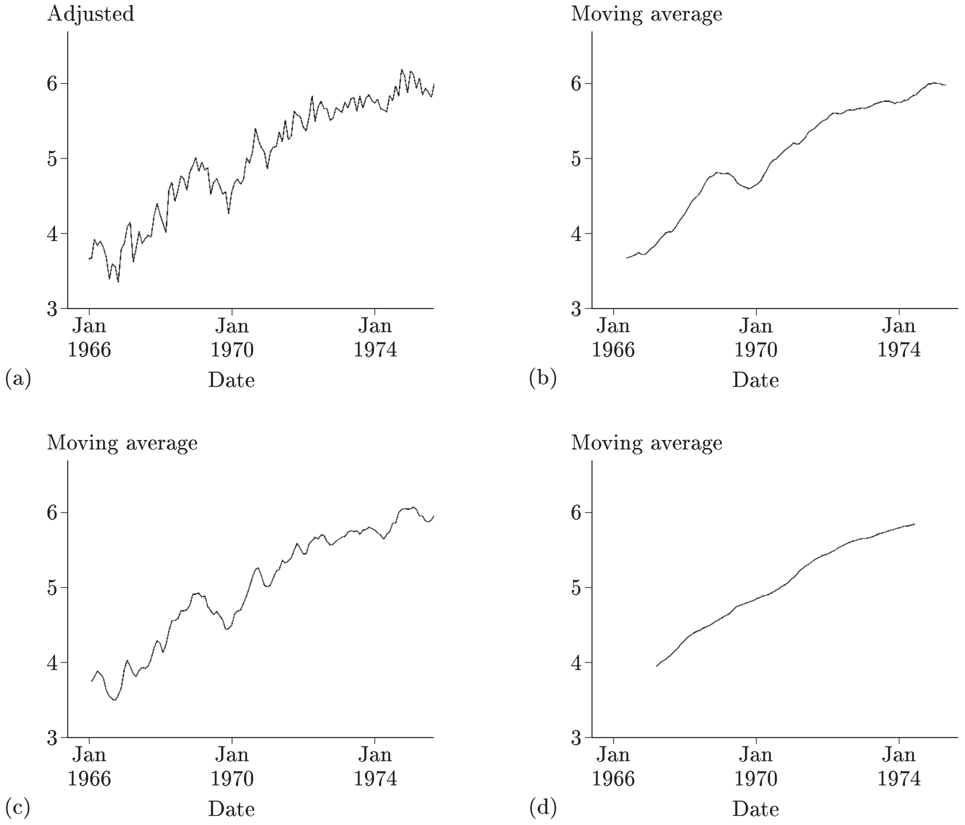

Figure 3

(a) The three simple moving averages are of orders 3, 11 and 31. Identify which subplot in Figure 3 corresponds to which moving average, and hence complete Table 6. Explain your answer. [*2*]

Table 6

Order	3	11	31
Subplot			

(b) In your view, which of the three moving averages provides the best estimate of the trend component of this time series? Justify your answer. [2]

Question 11

Figure 4(a) shows the monthly time series of armed robberies in Boston, from January 1966 to October 1975. Figure 4(b) shows the time series of logarithms of armed robberies.

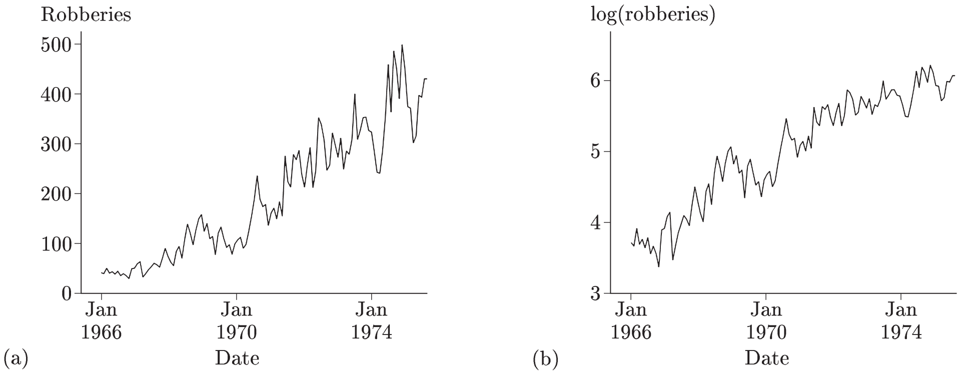

Figure 4

Explain whether an additive decomposition model, a multiplicative decomposition model or neither is appropriate for the untransformed data. [2]

Question 12

The estimated seasonal factors for a quarterly time series are as shown in Table 7.

Table 7

Quarter	Seasonal factor
1	-1.1634
2	?
3	1.0654
4	0.5418

(a) Obtain the seasonal factor for Quarter 2. [*1*]

(b) Briefly describe the seasonal variation in this time series. [*1*]

Question 13

Table 8 shows the first three values of a time series.

Table 8

t	x_t
1	0.52
2	0.36
3	0.48

(a) Use simple exponential smoothing with $\alpha = 0.75$ and starting value $\widehat{x}_1 = x_1$ to obtain the 1-step ahead forecast of X_4. *[2]*

(b) The true value of X_4 turns out to be 0.42. Calculate the SSE. *[2]*

Question 14

Figure 5 shows the time plot of monthly sales of a product (in £000).

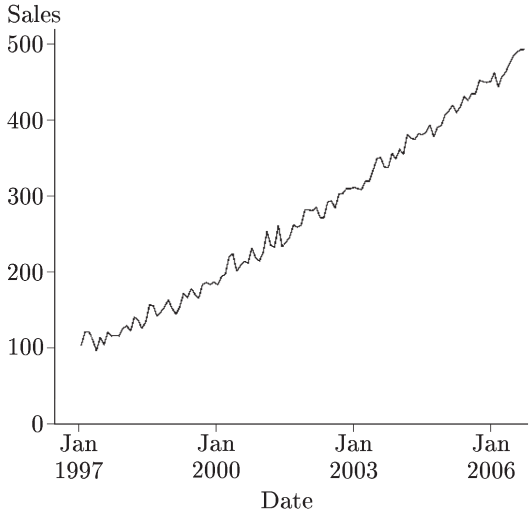

Figure 5

Suppose that the time series is not seasonal.

(a) Give two reasons why it would be more appropriate to use Holt's exponential smoothing rather than simple exponential smoothing or the Holt–Winters method to forecast future values of this time series. [*1*]

(b) For a given set of starting values, the optimal values for the smoothing parameters were found to be $\alpha = 0.81$, $\gamma = 0.09$. Briefly interpret these values in terms of dependence on observations in the recent and distant past. [*1*]

Question 15

An exponential smoothing method is used to obtain 1-step ahead forecasts for a time series. Figure 6 shows the correlogram of the forecast errors. The p value for the Ljung–Box test for zero autocorrelation at lags 1 to 20 of the forecast errors is 0.066.

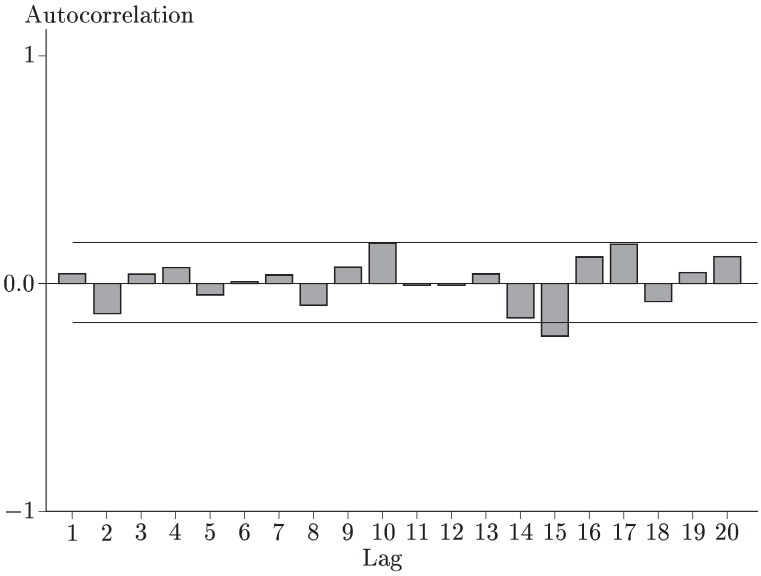

Figure 6

(a) Is it plausible that the autocorrelations at lags 1 to 20 are all zero? Give one reason in favour, and one against. [*2*]

(b) The time series comprises $n = 118$ observations. The sum of squared errors is 15 212.37, and the forecast for the next time point is $\widehat{x}_{119} = 499.78$. Obtain an approximate 95% prediction interval for X_{119}. (You may assume that any conditions required are satisfied.) [*2*]

Question 16

Temperatures of a chemical process (in °C) were measured at hourly intervals. Figure 7(a) shows the time plot of the temperatures, and Figures 7(b) and 7(c) show time plots of the first and second differences, respectively.

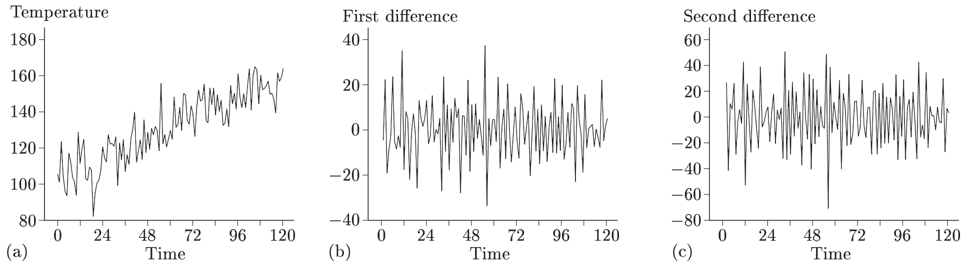

Figure 7

(a) Identify the order of differencing that appears to be required to obtain stationarity in mean. Justify your answer. [*2*]

(b) Does the differenced series that you identified in part (a) appear to be stationary in variance? Justify your answer. [*1*]

Question 17

Figure 8 shows the correlogram and partial correlogram of a stationary time series.

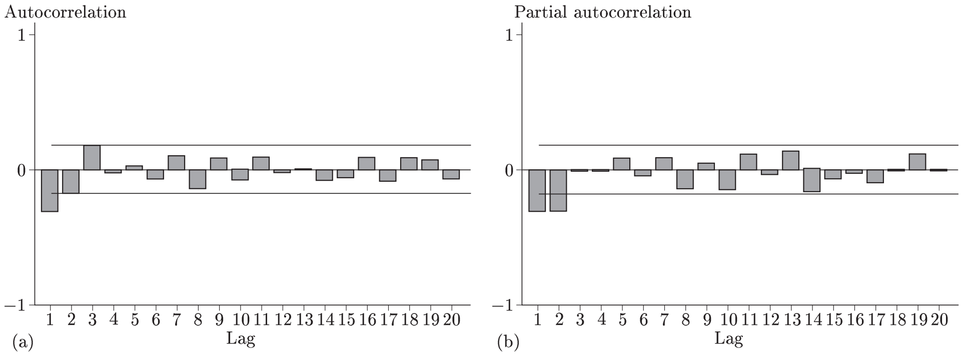

Figure 8

(a) Identify two plausible ARMA models for this time series. Justify your choices. [*4*]

(b) Apply the principle of parsimony to select one or more ‘best’ candidate models. [*1*]

Question 18

Let Y_t denote a non-seasonal time series, X_t a stationary series, and Z_t white noise. Classify each of the following models for Y_t as ARIMA(p, d, q) for suitable values of p, d and q. *[2]*

Model A: $X_t = 0.5X_{t-1} + Z_t - 0.2Z_{t-1}$ and $X_t = Y_t$.

Model B: $X_t = 0.05X_{t-1} - 0.4X_{t-2} + Z_t - 0.1Z_{t-1}$ and $X_t = Y_t - Y_{t-1}$.

For Examiner's use only:

Question No.	10	11	12	13	14	15	16	17	18	Total for Part 2
Mark										

PART 3

Questions 19 to 25 relate to Book 3 Multivariate analysis.

Write your answers in the spaces provided.

Question 19

Table 9 contains a small multivariate data set.

Table 9

X_1	X_2	X_3	X_4	X_5
3064	1201	10	361	2
3000	1053	11	338	3
3155	1133	19	393	2
3085	963	4	467	3
3245	1258	36	294	1
3045	1208	14	371	3

(a) State the number of variables and the number of observations in this data set. What are the values of x_{14} and x_{32}? [*1*]

(b) Calculate the covariance between X_1 and X_2. (You may assume that $\overline{x}_1 = 3099$ and $\overline{x}_2 = 1136$.) [*1*]

Question 20

A botanist measured various characteristics of plants from a particular species. She reduced the number of variables to be further analysed by using principal component analysis (PCA), applied to the data after the variables had been standardized. Some information about the variables is given in Table 10.

Table 10

Variable	Units of measurement	Mean	Standard deviation	Loadings of first principal component	Value for Plant A	Value of standardized variable for Plant A
Weight of first seed	mg	469	5.2	0.8	470	0.192
Age	days	153	4.6	0.1	150	−0.652
Height	cm	30.5	3.4	−0.3	35	?
Number of leaves		93	3.9	0.5	90	−0.769

(a) Give one reason why, in this case, it was important to standardize the data prior to using PCA. [*1*]

(b) Complete Table 10 by calculating the value of the standardized variable for height for Plant A. [*1*]

(c) Calculate the value of the first principal component for Plant A using the loadings given in Table 10. [*2*]

Question 21

The matrix scatterplot for some multivariate data is shown in Figure 9.

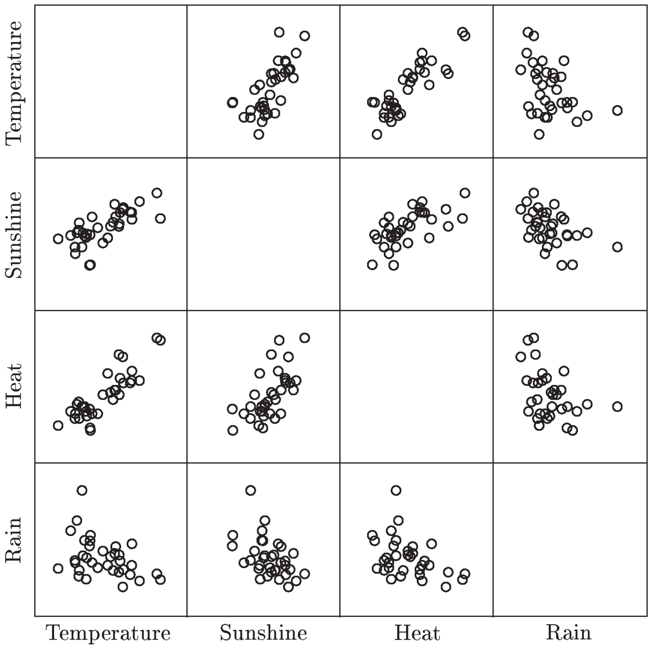

Figure 9

In this data set, the observations correspond to different calendar years, and the variables relate to various aspects of the weather at a particular location during the year.

(a) Use Figure 9 to comment on the relationships between the variables Temperature, Sunshine, Heat and Rain. *[1]*

(b) From Figure 9 it could be argued that the data contain a single outlier. Briefly describe in what way(s) this possible outlier is unusual. Is this outlying observation similar to the other observations in any way? If so, briefly describe in what way(s). *[2]*

Question 22

Five variables, $Y_1, Y_2, \ldots, Y_5$, were recorded for a set of 20 observations. The sample correlation matrix, calculated using these data, is

$$\begin{pmatrix} 1 & 0.77 & 0.09 & -0.74 & -0.70 \\ 0.77 & 1 & 0.13 & -0.80 & -0.76 \\ 0.09 & 0.13 & 1 & -0.17 & -0.23 \\ -0.74 & -0.80 & -0.17 & 1 & 0.78 \\ -0.70 & -0.76 & -0.23 & 0.78 & 1 \end{pmatrix}.$$

Use this correlation matrix to answer the following questions.

(a) Between which pair of variables does there appear to be the strongest linear relationship? Between which pair of variables does there appear to be the weakest linear relationship? [*1*]

(b) Describe the relationships between the variables implied by this correlation matrix. [*3*]

Question 23

A principal component analysis was carried out on a standardized data set consisting of 96 observations and 6 variables. The resulting scree plot is shown in Figure 10.

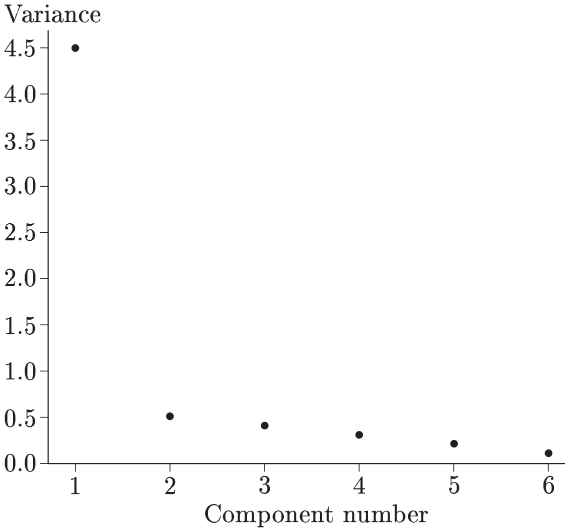

Figure 10

(a) Based on each of the following, how many components should be retained? In each case, explain your answer.

(i) The scree plot given in Figure 10. [*2*]

(ii) Kaiser's criterion. [*2*]

(b) Use Figure 10 to estimate the percentage variance explained by the first component. [*2*]

Question 24

In a study of cola drinks, 34 different varieties were tested. For each variety, 26 variables were recorded, including the calories per 100 ml, the price per 2-litre bottle and the pH. Also recorded for each variety was the conclusion of a panel of taste testers, categorized as very nice taste, average taste or worse than average taste.

(a) What is the maximum number of discriminant functions that are useful for discriminating between the conclusions of the taste-tester panel using the recorded variables? *[1]*

(b) A scatterplot of the varieties of cola with respect to the first two discriminant functions is shown in Figure 11.

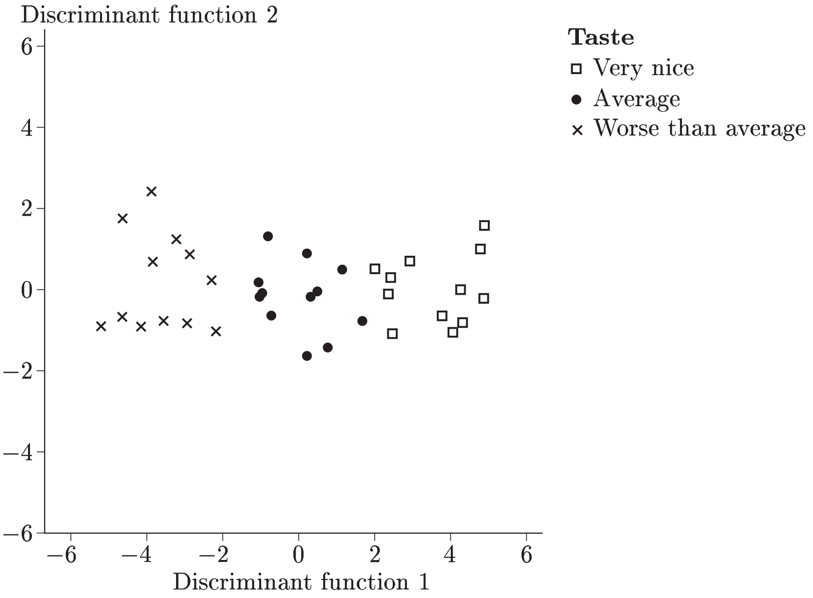

Figure 11

Briefly comment on the usefulness of the first two discriminant functions for distinguishing between the conclusions of the panel of taste testers. *[2]*

(c) For the first discriminant function, very nice tasting cola varieties had a mean score of 3.60, average tasting cola varieties had a mean score of 0.02, and worse than average tasting colas had a mean score of -3.62. Find the allocation rule for colas that have worse than average taste. (You may assume equal prior probabilities and equal costs of misclassification.) [*2*]

Question 25

In a principal component analysis of a four-dimensional data set based on standardized variables, two principal components are extracted. The loadings for the two principal components are given in Table 11.

Table 11

	Component 1	Component 2
Z_1	0.524	0.008
Z_2	-0.488	0.005
Z_3	0.510	0.682
Z_4	-0.477	0.732

Briefly interpret each of the components. [*4*]

For Examiner's use only:

Question No.	19	20	21	22	23	24	25	Total for Part 3
Mark								

PART 4

Questions 26 to 34 relate to Book 4 Bayesian statistics.

Write your answers in the spaces provided.

Question 26

In a study of immunity in Mexico, 86.5% of healthcare workers were found to have antibodies to rubella. 41.6% of the healthcare workers who had antibodies to rubella said that they had been exposed to rubella in the past, whereas 40.4% of all healthcare workers (whether or not they had antibodies to rubella) said that they had been exposed to rubella in the past.

Using this information, calculate the probability that a healthcare worker who says that he or she has been exposed to rubella has antibodies to rubella. [*2*]

Question 27

Suppose that X, the number of damaged items found in a packing case containing 100 items, has a binomial distribution $B(100, \theta)$, and that a prior for θ is $\text{Beta}(5, 3)$.

(a) Calculate the prior mode and variance. [*2*]

(b) A packing case is opened and found to contain 12 damaged items. State the posterior for θ, giving the values of any parameters. [*1*]

(c) The posterior mode and variance are 0.151 and 0.0012, respectively. How have the data affected the prior beliefs about θ? [*1*]

Question 28

Karen wishes to specify a prior distribution that summarizes her opinions about how late the bus she intends catching this evening is going to be. Let θ be the length of time (in minutes) that the bus is late. Karen believes that the bus will never be early and that the bus is most likely to be 5 minutes late.

(a) Suggest a suitable standard prior for the length of time that the bus is late. (You do not need to specify values for the parameters of the prior in this part of the question.) [*1*]

(b) Assuming that the second parameter, b, of Karen's prior is 2, calculate the value of parameter a for the prior distribution that you suggested in part (a). [*1*]

Question 29

Figure 12 contains plots of three prior distributions.

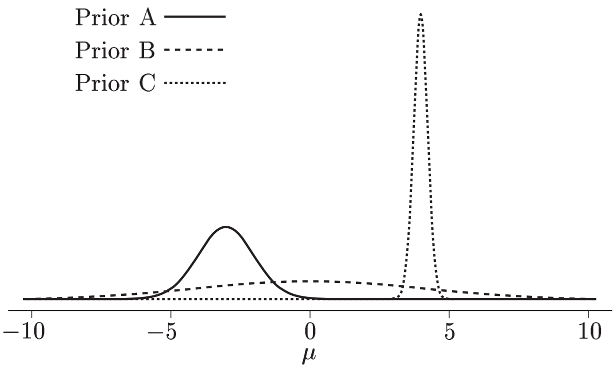

Figure 12

Place these prior distributions in order from weakest to strongest. [*1*]

Question 30

Suppose that some data x_1, x_2, x_3, x_4 can be modelled by a normal distribution, $N(\mu, 100)$. John believes that the median of μ is -2.4, and his assessed lower quartile and upper quartile for μ are -9.6 and 4.8, respectively.

(a) Calculate the values of the parameters a and b of the normal distribution $N(a, b)$ that best matches John's beliefs about μ. *[3]*

(b) As far as you can tell, is the prior that you specified in part (a) likely to be a good representation of John's beliefs? Why or why not? *[1]*

(c) Table 12 gives the observed values of x_1, x_2, x_3, x_4.

Table 12

i	1	2	3	4
x_i	8	-9	3	6

Using the prior that you specified in part (a) and the data in Table 12, calculate the posterior for μ. *[3]*

Question 31

In a study of businesses, last year's net profit X is to be collected for a sample of firms. X is to be modelled using a normal distribution $N(\mu, \sigma^2)$, where σ^2 is known. (A negative value of X indicates that the firm made a loss last year.) The aim of the study is to estimate μ, the mean net profit last year across all businesses.

(a) The researcher claims to have no idea whatsoever about the value of μ before the data are collected. Write down a prior for μ that represents this belief. [*1*]

(b) In the sample of businesses studied, the mean of last year's net profit turns out to be £10 124. So for this researcher's posterior for μ, state which of the following options is the correct one. Justify your answer. [*2*]

(i) The posterior mean of μ is greater than £10 124.

(ii) The posterior mean of μ is exactly equal to £10 124.

(iii) The posterior mean of μ is less than £10 124.

(iv) There is not enough information to say what the posterior mean of μ is.

Question 32

WinBUGS was used to sample from the posterior for a parameter μ. Two chains of sampled values, each of 1000 iterations, were obtained. A trace plot of the two chains superimposed on the same diagram is shown in Figure 13.

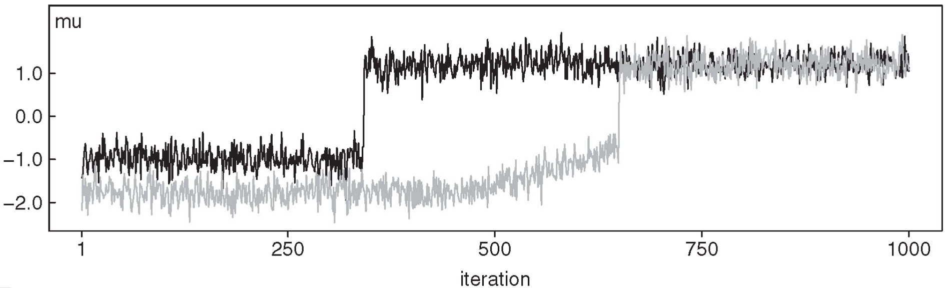

Figure 13

(a) Use Figure 13 to suggest a suitable burn-in. Explain your choice. [*2*]

(b) After the burn-in was discarded, the remaining values were analysed. A plot was produced of the estimated posterior density, together with the prior density and the likelihood. Give a reason why Figure 14 *cannot* be a copy of this plot. [*1*]

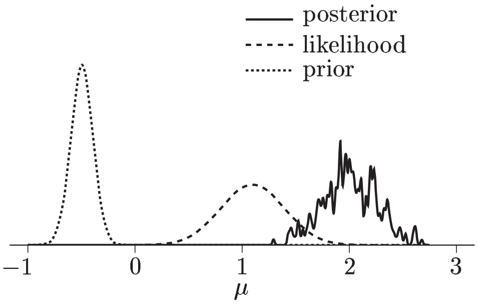

Figure 14

Question 33

A set of data is analysed independently by two friends, Jan and Chris. Using WinBUGS, Jan estimates that the 95% posterior credible interval for a parameter α is $(42.76, 54.14)$. Chris also uses WinBUGS and estimates that the 95% posterior credible interval for α is $(45.17, 56.76)$. Give two reasons why it is possible for both Jan and Chris to have estimated the posterior credible interval for α correctly. [*2*]

Question 34

Samples from the posterior distribution for three parameters, μ, σ and θ, were obtained using WinBUGS. Using these samples, the following numerical summaries were produced.

```
 node    mean      sd  MC error     2.5%   median   97.5%   start   sample
   mu   1.043  0.5681  0.005948   0.6471   0.9991    1.57       1    10000
sigma    20.0  0.2584  0.002743    19.56     20.0   20.45       1    10000
theta   50.12   28.98    0.2954    2.567    50.44   97.55       1    10000
```

(a) Write down an estimate for the posterior mean of μ and an estimate of the 95% posterior equal-tailed credible interval for μ. [*1*]

(b) Was the simulation run for long enough? Explain your answer. [*2*]

(c) Part of the model definition written in the BUGS language is as follows.

```
 model
{
    theta ~ dnorm(50, prec)
    prec <- 100
}
```

What is the distribution of θ according to this part of the model? Write down the mean and variance of θ. [*1*]

For Examiner's use only:

Question No.	26	27	28	29	30	31	32	33	34	Total for Part 4
Mark										

[END OF QUESTION PAPER]